DOCUMENT RESUME

ED 347 194 TM 018 687

AUTHOR Hecht, Jeffrey B.
TITLE Continuing Perspectives on the Analysis of Regression Residuals: Advocacy for the Use of a "Totter Index."
PUB DATE Apr 92
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); Graphs; *Least Squares Statistics; Mathematical Models; *Regression (Statistics); *Statistical Distributions; Theories
IDENTIFIERS DRAWREG Computer Program; *Outliers; *Residuals (Statistics); Totter Index

ABSTRACT

Previous research has demonstrated particular inadequacies in conventional methods used to identify outlier cases in bivariate regression models. Only through a combination of methods can one detect all of the deviant points potentially overinfluencing a regression model's parameters. This paper investigates whether a range of data points might exist that could be influential to a regression estimate, yet remain undetected by current methods. A computer program written for prior research (DRAWREG) was expanded to provide graphical displays of rejection boundaries according to the standardized, studentized, leverage, and Cook's Distance methods. Overlapping these regions revealed four geometric areas where influential points might exist, yet remain outside the usual detection frames. The traditional homogeneity assumption underlying least-squares regression (a rectangular pattern) did not agree with the usual graphic representation of bivariate relationships (an ellipsoid). The DRAWREG program was enhanced to compute and display an ellipsoid around the regression line and perpendicular coordinates. While further work is needed to clarify recommended boundaries for this region, it has potential as a single, adequate measure for influential point detection. Ten figures illustrate the discussion. A 12-item list of references and an appendix listing formulas are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JEFFREY B. HECHT
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED347194

Continuing Perspectives on the Analysis of Regression Residuals: Advocacy for the Use of a 'Totter Index'

by

Jeffrey B. Hecht
Department of Educational Administration and Foundations
College of Education
Illinois State University

This paper was presented at the Annual Meeting
of the
American Educational Research Association
April 20-24, 1992
San Francisco, California

TM018687

BEST COPY AVAILABLE

Abstract

In a previous paper (Hecht, 1991) I demonstrated particular inadequacies in conventional methods used to identify outlier cases in bivariate regression models. These methods consider a single case, by its very existence in a low probability geometric space, to be somehow deviant and therefore eligible for recoding or elimination. Consideration of distribution theory, however, leads to the conclusion that no matter how improbable, a single case must be allowed to exist anywhere in the bivariate space. It is the accumulation of many such deviant cases that should be cause to suspect underlying distributional assumptions. Instead, the criteria for single point outlier declaration should be that point's influence on an otherwise acceptable regression model. The prior work went on to demonstrate how existing methods of outlier detection are each particularly sensitive to the outlier's position relative to the bivariate mean ($\bar{X}$ and $\bar{Y}$) and to the regression line ($\hat{Y}$). It is only through a combination of methods that one can be assured of detecting all of the deviant points potentially overinfluencing a regression model's parameters.

This paper continues where the first left off. Intrigued with the inadequacies of current outlier detection methods I began to investigate whether there might exist a range of data points that could be influential to a regression estimate yet remain undetected by current methods. A computer program written for the prior research (DRAWREG) was expanded to provide graphical displays of rejection boundaries according to the standardized, studentized, leverage, and Cook's Distance methods. Overlapping these regions revealed four geometric areas where influential points might exist yet remain outside of these usual detection frames. At the same time I observed that traditional wisdom in regard to the homogeneity assumption underlying linear least-squares regression (a rectangular pattern) did not agree with the usually graphical presentation of bivariate relationships (an ellipsoid pattern). A search of statistical textbooks confirmed this approach, and while Afifi and Clark (1984) previously recognized this as an "ellipse of concentration" nothing could be found that converts this knowledge into a useful measure suitable for outlier detection.

Following this train of thought the DRAWREG program was enhanced to compute and display an ellipsoid around the regression line (X') and perpendicular (Y') coordinates. An ellipse can be constructed that represents the centroid of the bivariate data set by setting the center set at ($\bar{X}$, $\bar{Y}$), the major foci at predetermined *Z* values along X' and the minor foci at a ratio of the major While further work is needed to clarify recommended boundaries for this region, it has promise for use as a single, adequate measure for influential point detection.

Continuing Perspectives on the Analysis of Regression Residuals: Advocacy for the Use of a 'Totter Index'

Outlier data points are extreme cases that, in the opinion of each individual researcher, are not representative of the distribution of the true underlying data. The detection of outliers for consideration of inclusion 'as-is', recoding to a less extreme value, or outright elimination is an important concern. Regression models constructed that contain one or more extremely deviant points may over-represent error variance, significantly alter regression parameters (β_0 and β_1 in the two variable model), or both. When extreme, these outlier points can alter the predictive abilities of a potential model, reducing its usefulness to its real-world application (see Glenberg (1988, p. 439) for a simple yet convincing illustration of this point).

Unfortunately, current methods for detecting outlier data points prove to have certain drawbacks. Prior work (Hecht, 1991) examined four of the most commonly used methods: Standardized residuals, Studentized residuals, Leveraged residuals, and Cook's Distance (related to Mahalanobis Distance for N > 50 (Stevens, 1986)). An interesting anomaly was uncovered during this analysis, whereby many authors of statistical textbooks described the need for adhering to particular distributional assumptions during the construction of the regression model. These same authors would continue, however, to discuss outlier data points as if they should not exist at all due to the low probability of expected occurrence at a particular place in the assumed distribution. This prior work discussed that the bivariate normal distributional assumption allows for a data point to exist anywhere in the two dimension space. It is the probability of a particular density of points that is at issue, not the existence of a single point itself. Shifting the focus to the influence a single (or collection) of data points might have on regression model parameters allows for a more theoretically sound investigation.

Examination of these four methods revealed weaknesses in each as they were used to detect points more and more deviant through the entirety of the bivariate space. The method of Standardized and Studentized residual analysis best detected points deviant from the modeled regression line yet parallel to that line, without consideration of their overall distance from the bivariate mean ($\bar{X}$, $\bar{Y}$) along the regression line (see Appendix A for formulas). The difference between these methods becomes negligible when large N are considered. A large enough sample (N = 500) was employed throughout this investigation, so both methods will be presented simultaneously. Figure 1 displays a sample bivariate data set with a moderate correlation (r = .67) showing the regression line and both standardized rejection bands. A total of sixteen potential outlier points lie above and below these bands.

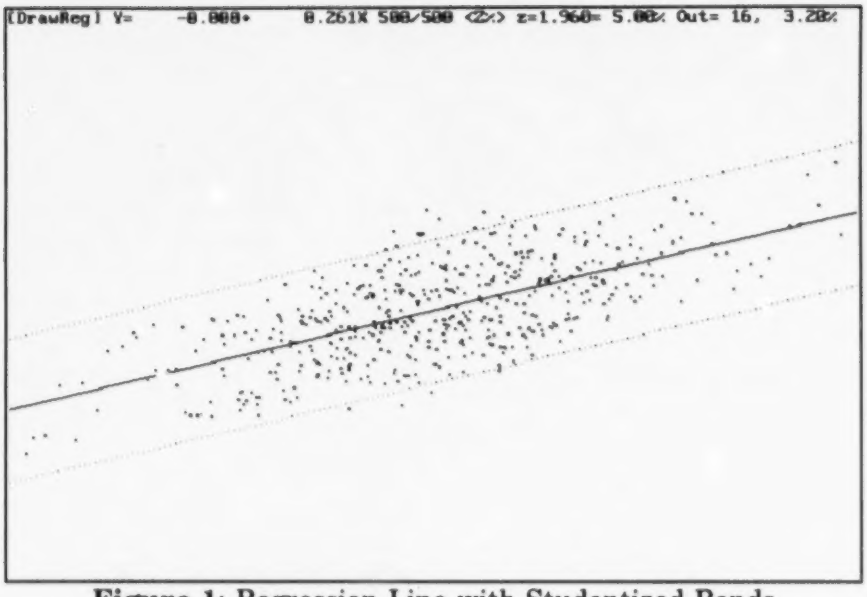

Figure 1: Regression Line with Studentized Bands

Leverage analysis is particularly suited to the detection of points deviant along the X-axis (see Appendix A). Setting a threshold level from $\bar{X}$ allows for the production of bands discriminating these points (Figure 2). In this case 42 points were identified using the typical critical value for leverage. Unfortunately, leverage bands do little to detect points deviant from the regression line or along the Y-axis.

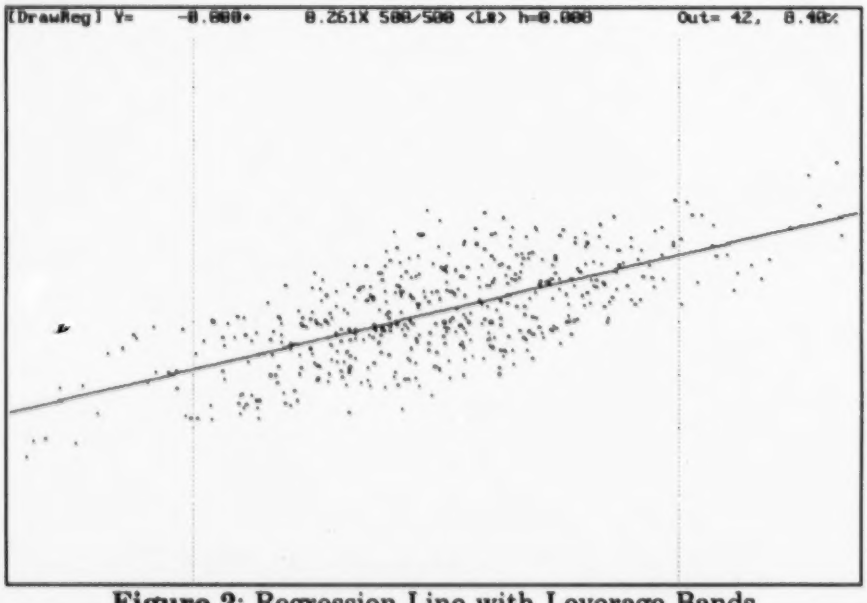

Figure 2: Regression Line with Leverage Bands

Cook's Distance represents a compromise in the need to detect outliers along both axes (See Appendix A). It is sensitive to the combination of studentized and leveraged distances, and produces a pair of varying pattern rejection bands depending on the X and Y values of each given data point. Figure 3 displays the same data set with the regression line and bands for Cook's Distance.

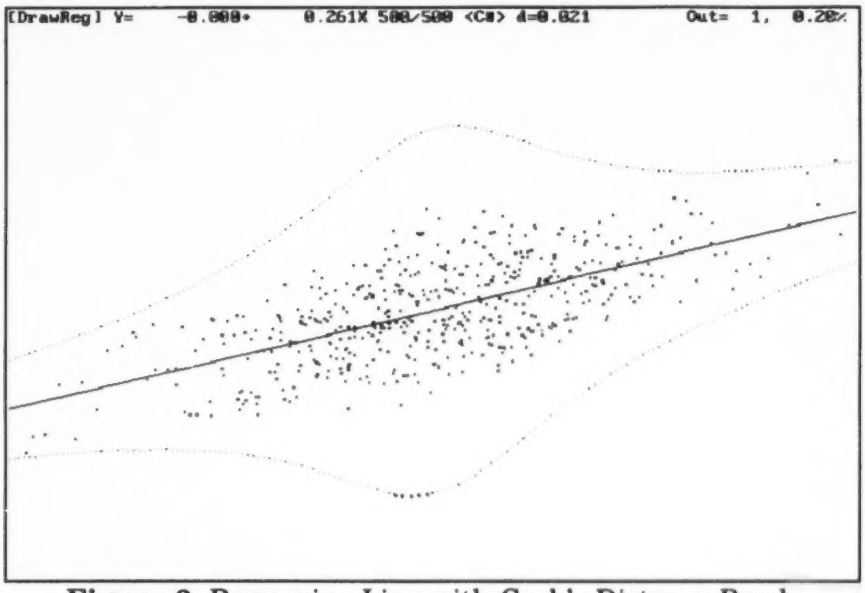

Figure 3: Regression Line with Cook's Distance Bands

An interesting side note needs to be made at this point. While examining the numerous data sets used in this research it became clear that the recommended critical value for Cook's Distance was usually too large for effective rejection ranges. Cook and Wiesberg, as described in Stevens (1986), "indicate that a [Cook's Distance] value of about 1 would generally be considered large" (p. 94). The recommended value meant, for most of the trial cases, that no cases at all would be identified as too deviant. Figure 3 displays bands using a critical Cook's distance of 0.021, just 0.001 smaller than the largest Cook's residual for this particular data set. A value of 1.000 (the recommended value) would not only fail to detect any outlier data points in this displayed data set, but would also draw band representations much larger -- and different looking -- than those displayed above. The smaller critical bands width is displayed in order to meaningfully convey the geometric shape of the rejection boundary.

Combining these methods together might seem like the safest approach to outlier detection. Indeed, the combination provides an overlap of X and Y (and X' and Y') coverage that would seem to include any potentially deviant points. Figure 4 demonstrates this combination.

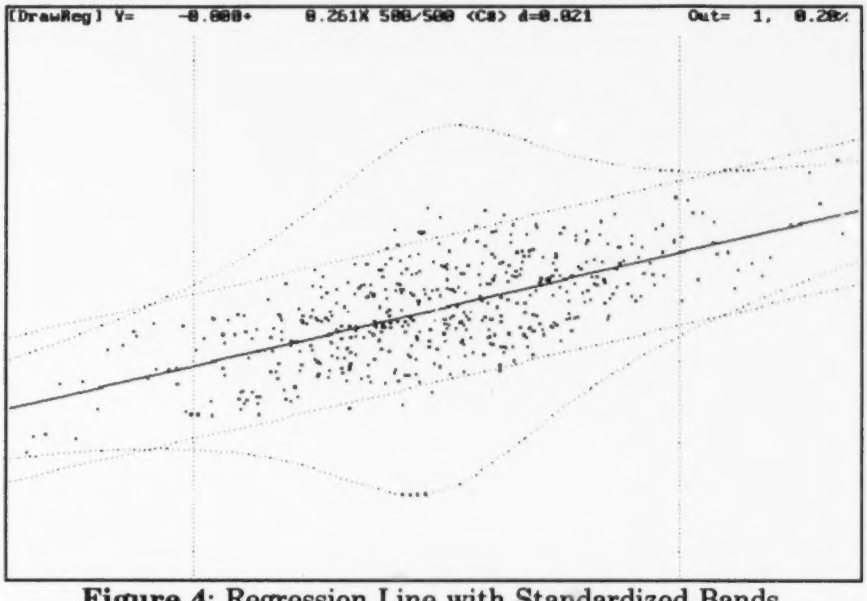

Figure 4: Regression Line with Standardized Bands, Leverage Bands, and Cook's Distance Bands

Even with the use of the aforementioned conservative critical value for the Cook's Distance Bands the combined area described is that of a trapezoid. Data points outside of this geometric boundary overly influence the regression model parameters creating inflated variances, shifted slope and/or intercept value, or both. Such points also have the potential for meaningfully influencing the correlation between the two data sets.

The question of what to do with potentially influential outlier data points goes beyond the scope of this research. Rather, the question for this study was whether this trapezoidal boundary adequately captures truly influential points across a wide range of data sets. If it did not, could there be another geometric shape that would both be easier to describe and be better able to identify the outlier points? This research investigated these two questions.

Method

500 paired points were randomly generated for bivariate data sets such that the bivariate means were 0.0 having approximately equal standard deviations ranging from 0.4 to 2.0 (in steps of 0.2, 9 steps total). The "X" and "Y" variables in these data sets were made to correlate between ±0.2 and ±0.8 (in steps of 0.05 for 26 total). A computer program was used to randomly select data points conforming to bivariate normal assumptions while meeting the variance and correlation parameters selected for each data set. This process resulted in 234 total different data sets being constructed.

Each of these data sets was input to the DRAWREG program. This program, written in QuickBASIC (Microsoft Corporation, 1989) for the PC/AT environment, calculated the regression parameters. The program was run on a PC/AT-486 computer running at 33Mhz with a Super-VGA monitor for display. After calculating the regression equation the data points, regression line, and residual analysis bands were displayed on the screen. The total number of data points identified by each method was noted, in addition to those points identified by more than one method. Notes were also made by the researcher as to general impressions concerning the shape of the distribution and position of potentially influential cases.

First Set of Results

Analysis of these 234 data sets confirmed the prior assumption of the utility of the different methods for outlier identification. No single method proved adequate for identifying all cases of outliers. The most promising method, Cook's Distance, was only useful for locating cases that were extremely deviant from the $(\bar{X}, \bar{Y})$ mean. Using the recommended critical value of 1.0 failed to identify many of the cases that clearly were not clustered with the remainder of the data set. While a tighter tolerance for C_d might correct for this, no natural value (or method of calculating such a value) presented itself. This remains a useful point for further research.

The combination of the Studentized bands and the Leverage bands, when both used together, seemed to afford the most reasonable approach. Both extreme X and Y outliers were detected as lying outside of the trapezoidal shape described by these bands. This is not unexpected, as the assumption of homoscedasticity under which least-squares linear regression operates calls for a rectangular kind of distribution along the length of the regression line. Ott (1988), as well as many others (Hays, 1988; Minium, 1978; and Harnett, 1982) make this point quite plain in their texts (see Figure 5 from Ott, p. 339).

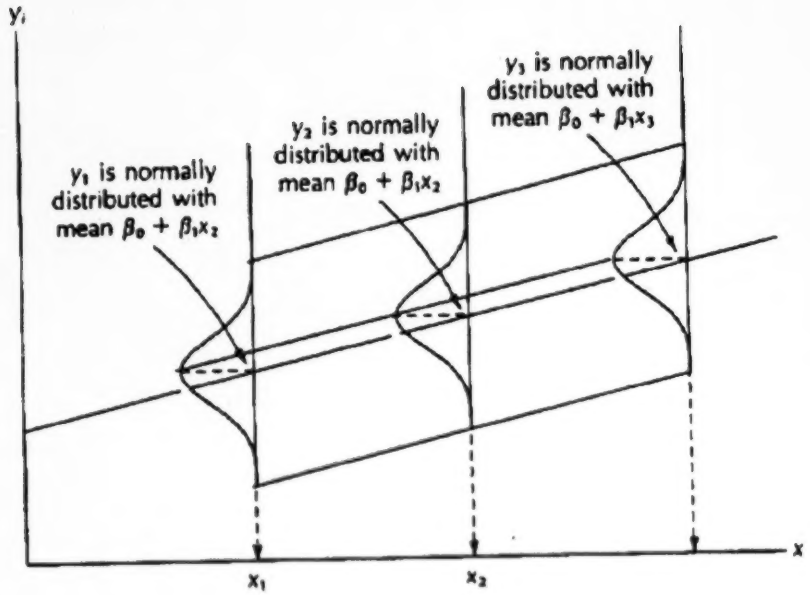
y_i
y_3 is normally distributed with mean $\beta_0 + \beta_1 x_3$
y_2 is normally distributed with mean $\beta_0 + \beta_1 x_2$
y_1 is normally distributed with mean $\beta_0 + \beta_1 x_1$
x
x_1
x_2

Figure 5: Assumption of Homoscedasticity

Draper (1981) discussed this in a similar fashion, without operationalizing to the Studentized and Leverage bands, by reasoning a "joint 95% confidence region for the true parameters, β_1 and β_2" (p. 95). He further reasoned that, while this trapezoid might serve as an adequate estimator for this "joint confidence region", an ellipse would do a much better job. Such a "confidence ellipse", as he called it, would not be difficult to construct and would do a much better job defining the boundaries than any straight lines (in a trapezoidal shape) would do.

Understanding why this might be so requires a little visualization. The normative bivariate data set with X and Y both following a normal distribution, plotted on a similar mean (or standardized), with X and Y uncorrelated ($r_{XY} = 0$) takes on the familiar dome shaped structure (see Figure 6 from Glass & Smith, 1984). It has the greatest frequency of scores near the common mean ($\bar{X}$, $\bar{Y}$), and describes a circle at its base. While the base is actually asymptotic to the XY plane, distributional theory states that it should radiate equally in all directions; that is, the decreasing probability of point density is equal all the way around its circular base. In other words most points would be found clustered around the bivariate mean, with fewer and fewer points (occurring equally in all directions) as one moves away from the joint mean.

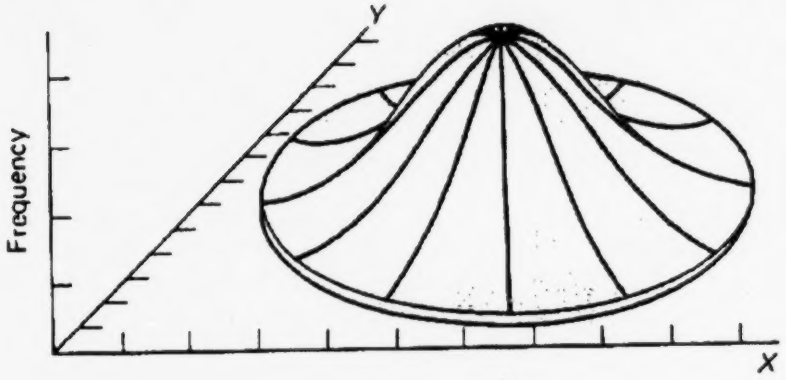

Figure 6: The Bivariate Normal Distribution, $r_{XY} = 0$

When X and Y follow the normal distributional assumptions and X is correlated to some degree with Y, the bivariate frequency distribution shape changes. As the correlation increases in magnitude the base becomes more elliptical and the distribution of the Y variables, though still normal with respect to balanced variability, increase in kurtosis (see Figure 7 from Kleinbaum, Kupper, & Muller, 1988). The greater the correlation, the more sharply defined the ellipse and the taller each normally distributed cross segment of Y until finally, at $|r| = 1$ the ellipse is complete subsumed into a line with normal density distribution along its X' length.

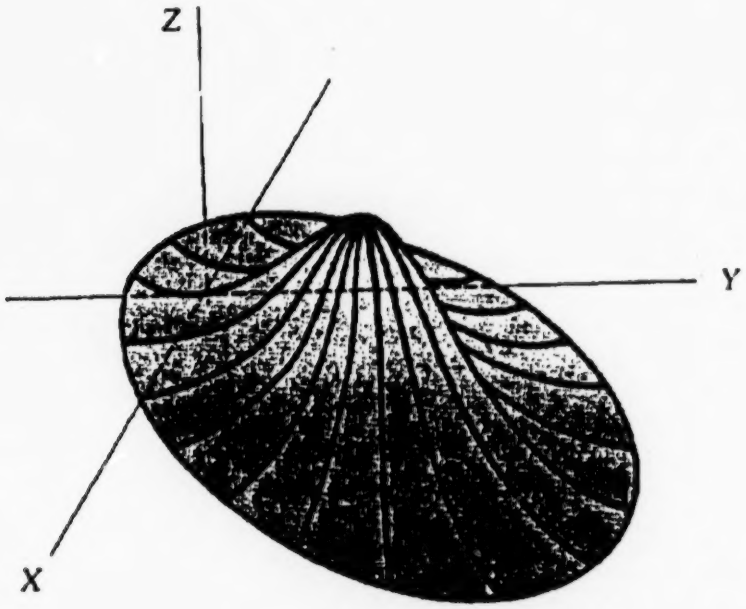

Figure 7: Bivariate Distribution, X Correlated with Y

Afifi and Clark (1984) also discussed this phenomenon, calling it the "ellipse of concentration" (p. 93). Figure 8 (from p. 95 in their text) pictures this relationship, which they explain as follows:

> For ellipses of concentration the center is at the point defined by the means of X and Y. The directions and lengths of the major and minor axes are determined by the two variances and the correlation coefficient. For fixed values of the variances the ratio of the length of the minor axis to that of the major axis, and hence the shape of the ellipse, is determined by the correlation coefficient ρ. (p. 94)

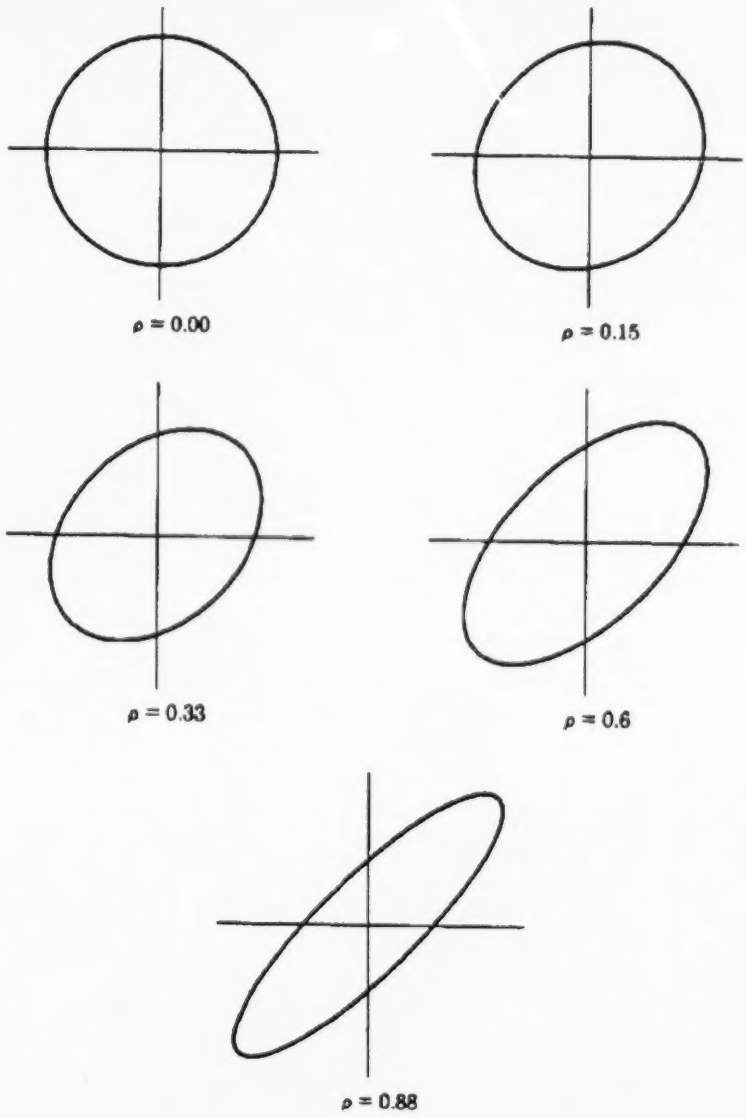

Figure 8: Ellipses of Concentration for Various Correlations

From these arguments it would appear than an ellipse, centered on the bivariate mean and of the proper length on both axes, would serve as a better discriminator for identifying outlying cases. This ellipse would conform to the planar correlated bivariate distributional shape. Under the trapezoidal view it would be assumed that the data modeling the regression line is just a subset of a larger set essentially extending out to infinity off both ends of the predicted line. The standardized or studentized bands are the only criteria applying under that perspective, as they do not concern themselves with deviant X cases and allow for homoscedasticity along the entire length of the predicted line (and beyond).

The correlated bivariate assumption, however, can have definite boundaries on both the X' and Y' axes according to the shape of the ellipse. Homoscedasticity is preserved through a decrease in the proportion of cases the further one moves from the bivariate mean. While an elliptical shape might adequately capture the essence of this assumption, no statistic exists for calculating or displaying such a curve, or for determining whether points are deviant based on this assumption.

In an attempt to enlighten this issue, and to possibly uncover a reasonable means for using this Ellipse of Concentration as a single measure for bivariate outlier detection, the data sets used in the first part of this study were examined for a second time. First, the DRAWREG program was enhanced to allow for the display of an ellipse shape centered on ($\bar{X}$, $\bar{Y}$) (see Figure 9). The major axis of the ellipse was placed along the regression line (X'), with the major foci initially position at ±2.0 standard deviations from $\bar{X}$. The minor axis was aligned on the perpendicular to the regression line (Y').

Several different values were used to establish the placement of the minor foci, with the program accepting a ratio value between the X' standard deviation and the Y' standard deviation as the critical value. When this ratio equaled its maximum of 1.0, the minor foci would be the same number of standard deviation units from $\bar{Y}$ as the major foci were from $\bar{X}$. This would describe a perfect circle, representing no correlation. Lesser values would make the shape more elliptical, representing increasing magnitudes of correlation. Finally, at a ratio of 0.0 the minor foci would collapse into $\bar{Y}$, rendering the ellipse as simply a straight line.

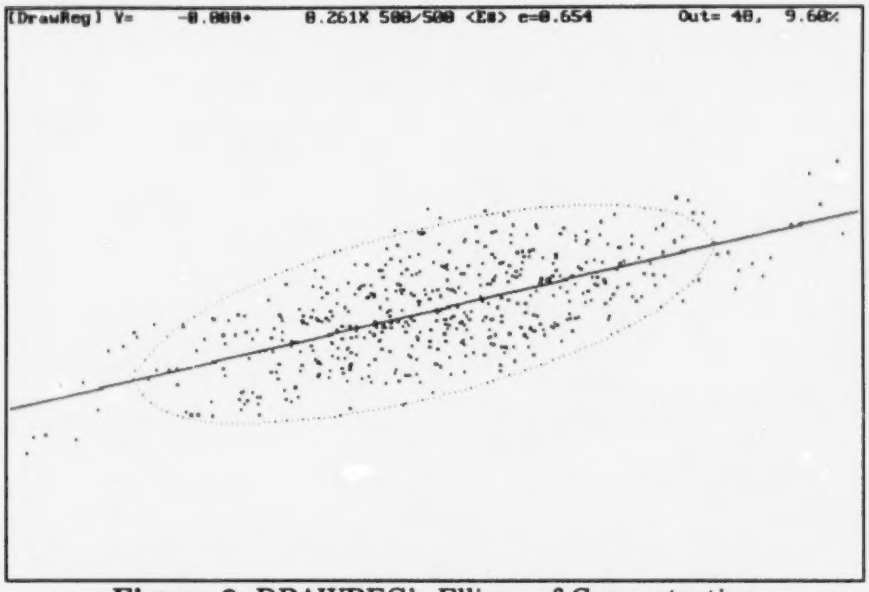

Figure 9: DRAWREG's Ellipse of Concentration

Each data set was examined in this manner, with the value for this ratio adjusted and noted. A balance point was sought where the ellipse would describe approximately the same interior space as that allowed by the trapezoid shape tendered by the Studentized and Leverage bands (see Figure 10).

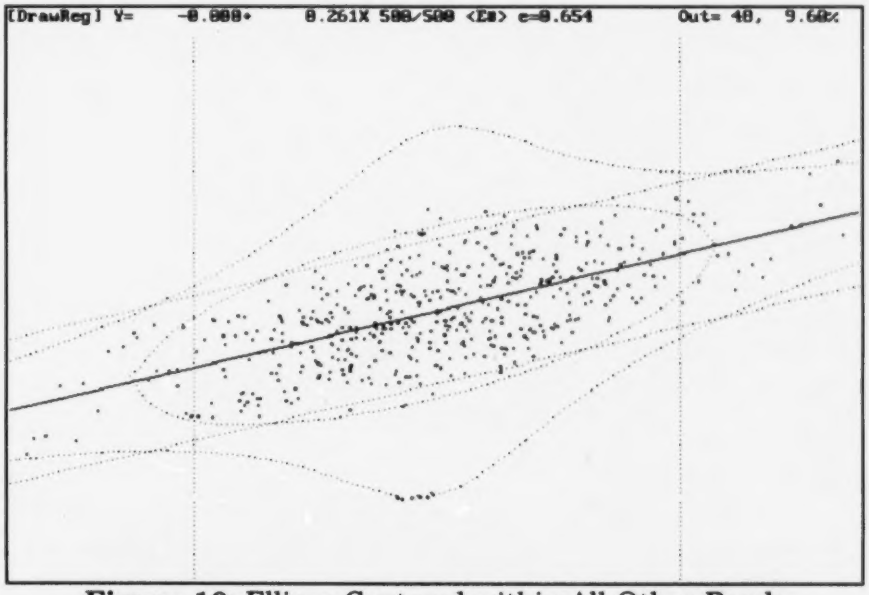

Figure 10: Ellipse Centered within All Other Bands

Second Set of Results

The ellipse of concentration appears to be a better choice for detecting potential outlier cases in the bivariate regression than any of the other methods mentioned. The ellipse identifies deviant cases along the X' axis (in a fashion similar to the Leverage bands) and along the Y' axis (like the Standardized and Studentized bands). Further, it does not suffer from the extreme liberalism of Cook's Distance. Finally, it represents a conceptually elegant and computationally simple single method, using the formula:

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1$$

where:

* the center of the ellipse is at $(\bar{X}, \bar{Y})$
* X' intercepts are located at **a** and **-a**
* Y' intercepts are located at **b** and **-b**

Experimentation with the values for **±a** and **±b** seemed to suggest that **a** might be initially set to ±1.96sd and **b** set to ± **a***(1 - |r|). The boundary of the ellipse can be calculated by solving the above formula for Y, then comparing each actual (X, Y) point with its calculated ellipse boundary. Those points having a true Y value greater than the boundary (above the line) or lesser (for below the line) would be suspect and considered outliers.

The ellipse produced by these values seemed to capture the inner area of the trapezoid while remaining true to the presumed shape of the correlated bivariate planar normal distribution. These corners represent places where an extreme data point could exist, having influence over the regression parameters but remaining undetected by the other methods. Additional experimentation and examination will have to be done to refine these suggested values for actual practice.

Conclusions

Detection of potentially influential, yet deviant, data points is a necessary step in the construction of reliable regression models. This paper demonstrates that neither the Standardized, Studentized, Leverage, or Cook's Distance methods adequately detect all cases of outliers. A combination of both the Studentized and Cook's Distance appears minimally necessary to insure that most deviant points are uncovered. Unfortunately, this dual methodology requires the computation and examination of two separate statistics. It is further complicated by the fact that it is only an approximation of the desired bivariate space, having four corner within its own trapezoidal structure where influential points may accumulate.

A critical region described by an ellipse would be more in line with the assumptions underlying the correlated bivariate data set. The use of an elliptical shape can be likened to the use of a "teeter-totter" by young children. It is less sensitive to heavier points closer to the middle than out along its extremes. Like a true teeter-totter, however, it does have certain maximum thresholds it will tolerate before tipping to one side or the other.

It is easy, however, to envision the creation of a statistic **e** representing this "Ellipse of Concentration". Statistic **e** might then be defined as:

$$e = \hat{Y}_X - \left(\beta_0 + \beta_1 * \pm\sqrt{b^2 * (a^2 - X^2)} \right)$$

A value for $|\hat{Y}| - |e| \leq 0$ would denote a point inside the ellipse. Values of $|\hat{Y}| - |e| > 0$ would identify an outlier and potentially influential point, with larger values of **e** indicating greater influence. While this paper suggests values for the construction of such an ellipse (**a** and **b** in the ellipse formula) additional research will need to be done to confirm and refine suggested critical values for this method. The simplicity of this design, its conceptual match to the bivariate distributional theory, and its potential ease of use all argue strongly for this continued investigation towards and eventual use of this 'totter index'.

References

Afifi, A. & Clark, V. (1984). *Computer-aided multivariate analysis*. Belmont, CA: Wadsworth.

Draper, N.R. & Smith, H. (1981). *Applied regression analysis*. New York, NY: John Wiley and Sons.

Glass, G. & Hopkins, K.D. (1984). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice Hall.

Glenberg, A.M. (1988). *Learning from data: An introduction to statistical reasoning*. San Diego, CA: Harcourt Brace Jovanovich.

Harnett, D.L. (1982). *Statistical methods*. New York, NY: Addison-Wesley.

Hays, W.L. (1988). *Statistics*. Orlando, FL: Holt, Rinehart and Winston.

Hecht, J.B. (1991, April). *Least squares linear regression and Schrodinger's Cat: Perspectives on the analysis of regression residuals.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. TM 016 495)

Kleinbaum, D.G., Kupper, L.L., & Muller, K.E. (1988). *Applied regression analysis and other multivariable methods.* Boston, MA: PWS Kent.

Microsoft Corporation (1989). *Microsoft QuickBASIC 4.5*. Redmond, WA: Author.

Minium, E.W. (1978). *Statistical reasoning in psychology and education*. New York, NY: John Wiley and Sons.

Ott, L. (1988). *An introduction to statistical methods and data analysis*. Boston, MA: PWS Kent.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Appendix A - Formulas

Standardized residual: $Z_{Y_{Res}} = \frac{Y_{Res}}{s_{Res}}$

where the critical z is: User selected (follows Z dist for n>50)

Studentized residual: $r_i = \frac{Y_i - \hat{Y}_i}{s\sqrt{1 - h_i}}$

where the critical r is: User selected (follows Z dist for n>50)

Leverage: $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n - 1)\, S_X^2}$

where the critical h is: $h_i > \frac{2\,(k + 1)}{n}$

[Note: Mahalanobis' Distance (D^2) was not used since D^2 is proportional to $(n-1)h_i$ when n>50]

Cook's Distance: $d_i = \frac{(Y_i - \hat{Y}_i)^2\, h_i}{(k + 1)\, s^2\, (1 - h_i)^2}$

where the critical d is: $d_i > 1$

Ellipse of Concentration: $e = \hat{Y}_X - \left(\beta_0 + \beta_1 * \pm\sqrt{b^2 * (a^2 - X^2)} \right)$

where the critical e is: $|\hat{Y}| - |e| > 0$